**Data Sets Description**

Real-Time Sensor Anomaly Detection and

Identification in Automated Vehicles

**1. General information**

This document provides an overview of the data sets used in the numerical experiments for the paper ‘Real-Time Sensor Anomaly Detection and Identification in Automated Vehicles’ published in the journal *IEEE Transactions on Intelligent Transportation Systems*. The data features extracted from the Safety Pilot Model Deployment (SPMD) dataset used in this study include the in-vehicle speed (denoted as sensor 1), GPS speed (sensor 2), and in-vehicle acceleration (sensor 3) for one of the test vehicles. As described in the paper, we consider the following four types of anomalies.

1) Instant: A sharp, unexplained change in the observed data between two successive sensor readings.

2) Constant: A temporarily constant observation that is different from the “normal” sensor readings and is uncorrelated to the underlying physical phenomena.

3) Gradual drift: A small and gradual drift in observed data during a time period. It can result in a large discrepancy between the observed data and the true state of the system in time.

4) Bias: A temporarily constant offset from the sensor readings.

In the following, details are provided regarding what data sets to use for specific experiments and how to generate these data sets using the code provided. The data set without any anomalies titled ‘Data without anomalies’ is included in the main folder.

**2. Single Anomaly Types**

In the ‘Single Anomaly Types’ folder, the train and test data sets for the numerical experiments in Section IV-A are provided. Specifically, the sensor values with anomalies and the ground truth (anomalous or normal) for each sensor are included. In addition, Python codes used to generate the data are provided. To generate the data sets, or to modify the experiments, simply change the duration/magnitude of a given experiment by following the comments in the notebooks. For the sensor values and ground truth files, the train set constitutes the first 23840 time epochs (80% of the data set). In addition, the test set constitutes the remaining data entries, i.e., the last 5960 time epochs in the respective files.

**3. Mixed Anomaly Types**

In the ‘Mixed Anomaly Types’ folder, the train and test data sets for one of the 10 iterations of the experiments in Section IV-B are provided. Similar to the data included in the ‘Single Anomaly Types’ folder, for the sensor values and ground truth files, the train set constitutes the first 23840 time epochs (80% of the data set) of each anomaly type. In addition, the test set constitutes the remaining data entries, i.e., the last 5960 time epochs for each anomaly type in the respective files. To generate multiple iterations of the same anomaly settings, simply change the number inside ‘np.random.seed(x)’ in the code. In addition, to generate anomalies at a rate of 1%, as opposed to 5%, simply change 0.05, 0.033, and 0.0167 in the code to 0.01, 0.0066, and 0.00334, respectively.